Automatic Image Annotation using Deep Learning and Fisher Vectors

Prof. Lior Wolf
The School of Computer Science, Tel-Aviv University

BEN KLEIN, GUY LEV, GIL SADEH, L. WOLF.
ASSOCIATING NEURAL WORD EMBEDDINGS WITH DEEP IMAGE REPRESENTATIONS USING FISHER VECTORS. CVPR, 2015
Future work: Automatic textural scene description

A camera is viewing a scene and outputs a textural description of the activity over time.

The engine learns from pairs of the form: image + caption (weak supervision).

One child is opening a cabinet while the other kid is talking on the phone.

The kid is checking the refrigerator.

The girl is operating the microwave.
Task I: Image Annotation

- A man is playing guitar
- Two girls are playing soccer
- A man is walking down the street
- A man is climbing a mountain
- A boy is riding a bicycle
Task I: Image Annotation

A man is playing guitar
Two girls are playing soccer
A man is walking down the street
A man is climbing a mountain
A boy is riding a bicycle
Task II: Image Search

A man is playing guitar
Task II: Image Search

A man is playing guitar
Task III: Description Synthesis

Input: new unseen image

Output: new description in English

Two girls playing soccer.
Agenda

Image representation

Text representation/ NLP

Linking the two

Synthesis
We employ a pretrained Deep Convolutional Neural Network (CNN)

VGG by Andrea Vedaldi and Andrew Zisserman
Text representation

NLP as computer vision guys

1. What are the local descriptors
2. How to combine (pool) the local descriptors
Word2Vec

Word2Vec transforms word in English to representation with a semantic properties.

\[
\text{\texttt{word2vec("A") = (131, 128, 111, 10, ..., 14, 11) \in R}^D
\]
\[
\text{\texttt{word2vec("playing") = (11, 61, 2, 13, ..., 11, 10) \in R}^D
\]
\[
\text{.}
\]
\[
\text{.}
\]
\[
\text{.}
\]
\[
\text{\texttt{word2vec("guitar") = (21, 122, 14, 1, ..., 110, 1) \in R}^D
\]

In the beginning God created the heaven and the earth. And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. And God said, Let there be light: and there was light. And God saw the light, that it was good: and God divided the light from the darkness.

Word2Vec Neural Network

Input Layer

Hidden Layer

Output Layer

Huffman code of ‘Spirit’
In [24]: model.most_similar('python')

Out[24]:
[['scripting', 0.912078857421875],
['bash', 0.9030072093009949],
['perl', 0.897027850151062],
['tcl', 0.8833462595939636],
['ruby', 0.8729183673858643],
['c++', 0.8634607195854187],
['jython', 0.8467384576797485],
['groovy', 0.846560001373291],
['lua', 0.8416544795036316]]
The traditional computer vision pipeline

Local Feature Extraction  Pooling  Classifier  “dog”
The traditional computer vision pipeline

\[(31, 128, 11, 0, \ldots, 4, 1) \in \mathbb{R}^d\]
The traditional computer vision pipeline

Local Feature Extraction

$(31, 128, 11, 0, \ldots, 4, 1) \in \mathbb{R}^d$

$(11, 61, 2, 3, \ldots, 11, 10) \in \mathbb{R}^d$
The traditional computer vision pipeline

Local Feature Extraction

\[(31, 128, 11, 0, ..., 4, 1) \in \mathbb{R}^d\]
\[(11, 61, 2, 3, ..., 11, 10) \in \mathbb{R}^d\]
\[\cdot\]
\[\cdot\]
\[(21, 22, 4, 1, ..., 110, 1) \in \mathbb{R}^d\]
The traditional computer vision pipeline

\[(31, 128, 11, 0, \ldots, 4, 1) \in \mathbb{R}^d\]
\[(11, 61, 2, 3, \ldots, 11, 10) \in \mathbb{R}^d\]
\[\vdots\]
\[\vdots\]
\[(21, 22, 4, 1, \ldots, 110, 1) \in \mathbb{R}^d\]

\[x = (0.12, -0.13, 0.14, \ldots, -0.1) \in \mathbb{R}^m\]
The traditional computer vision pipeline

\[ x = (0.12, -0.13, 0.14, \ldots, -0.1) \in \mathbb{R}^m \]

Classifier

\[
\begin{cases}
  w \cdot x \geq 0 & \text{"dog"} \\
  w \cdot x < 0 & \text{"cat"}
\end{cases}
\]
Pooling methods – Bag of words - NLP

$I("man" \in Document)$

$(0, 1, 0, 0, 1, 0, \ldots, 0, 1) \in \mathbb{R}^V$

$I("work" \in Document)$
Pooling methods – Bag of words - Vision

Local Feature Extraction

(31, 128, 11, 0, ..., 4, 1) ∈ R^d
(11, 61, 2, 3, ..., 11, 10) ∈ R^d
...
Pooling methods – Bag of words – Vision

(31, 128, 11, 0, ..., 4, 1) ∈ R^d
(11, 61, 2, 3, ..., 11, 10) ∈ R^d
.
.
(21, 22, 4, 1, ..., 110, 1) ∈ R^d
Pooling methods – Bag of words – Vision

\[(31, 128, 11, 0, \ldots, 4, 1) \in \mathbb{R}^d\]
\[(11, 61, 2, 3, \ldots, 11, 10) \in \mathbb{R}^d\]
\[\vdots\]
\[\vdots\]
\[(21, 22, 4, 1, \ldots, 110, 1) \in \mathbb{R}^d\]
Pooling methods – Bag of words – Vision

\[(31, 128, 11, 0, \ldots, 4, 1) \in \mathbb{R}^d\]
\[(11, 61, 2, 3, \ldots, 11, 10) \in \mathbb{R}^d\]
\[ \ldots \]
\[ (21, 22, 4, 1, \ldots, 110, 1) \in \mathbb{R}^d \]

\[ I(\text{blue cluster} \in \text{Image}) \]
\[ (0, 1, 0, 0, 1, 0, \ldots, 0, 1) \in \mathbb{R}^V \]

\[ I(\text{yellow cluster} \in \text{Image}) \]
Pooling methods – Bag of words – Vision

Problem: Losing information – “A nose is a nose”
Pooling methods – Fisher Vector

\[(31, 128, 11, 0, ..., 4, 1) \in \mathbb{R}^d\]
\[(11, 61, 2, 3, ..., 11, 10) \in \mathbb{R}^d\]
\[\vdots\]
\[\vdots\]
\[\vdots\]
\[(21, 22, 4, 1, ..., 110, 1) \in \mathbb{R}^d\]
Pooling methods – Fisher Vector

\[(31, 128, 11, 0, \ldots, 4, 1) \in \mathbb{R}^d\]
\[(11, 61, 2, 3, \ldots, 11, 10) \in \mathbb{R}^d\]
\[\ldots\]
\[(21, 22, 4, 1, \ldots, 110, 1) \in \mathbb{R}^d\]

Diagonal-Covariance Gaussian Mixture Model

F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. (CVPR 2007)
Pooling methods – Fisher Vector

Diagonal-Covariance
Gaussian Mixture Model

- Probabilistic Model
  - Defined by K Multivariate Gaussian with a Diagonal Covariance
- The parameters of the model are:
  - $\tau_1, ..., \tau_k \in R$
  - $\mu_1, ..., \mu_k \in R^d$
  - $\sigma_1, ..., \sigma_k \in R^d$
- The parameters are learned by the EM algorithm
Gaussian Mixture Model - EM

Let $X_{trn} = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^D$ be the train set for the EM.

EM algorithm for GMM:
- Randomly initialize all the parameters $\lambda^{(0)} = \{\mu^{(0)}, \sigma^0, \tau^0\}$
- For $t=0...T$ (Or until convergence):
  - Estimation Step:
    $$p(z_i = k| x = x_i; \lambda^t) = T_{k,i}^t = \frac{\tau_k^t g(x; \mu_k^t, \sigma_k^t)}{\sum_{r=1}^{K} \tau_r^t g(x; \mu_r^t, \sigma_r^t)}$$
  - Maximization Step:
    $$\tau_{k}^{(t+1)} = \frac{\sum_{i=1}^{N} T_{k,i}^{(t)}}{\sum_{r=1}^{K} \sum_{i=1}^{N} T_{r,i}^{(t)}}$$
    $$\mu_{k,d}^{(t+1)} = \frac{\sum_{i=1}^{N} T_{k,i}^{(t)} \cdot x_{i,d}}{\sum_{i=1}^{N} T_{k,i}^{(t)}}$$
    $$(\sigma_{k,d}^{(t+1)})^2 = \frac{\sum_{i=1}^{N} T_{k,i}^{(t)} (x_{i,d} - \mu_{k,d}^{(t+1)})^2}{\sum_{i=1}^{N} T_{k,i}^{(t)}}$$
Pooling methods – Fisher Vector

For a given GMM, and a set of vectors, \( X = \{x_1, \ldots, x_n\} \in \mathbb{R}^D \)
We can compute the log-likelihood of \( Y \) given the parameters of the model \( \lambda = \{\mu, \sigma, \tau\} \).

\[
L(\lambda|X) = \sum_{i=1}^{n} \log(p(x_i|\lambda))
\]

For each parameter \( \text{param} \in \lambda \), we can compute:

\[
\frac{\partial L(\lambda|X)}{\partial \text{param}}
\]

The vector \( \text{FV}(X) \) is the vector of all the gradients.
Pooling methods – Fisher Vector

Therefore: \( FV(X) \in R^{(2D+1) \cdot K} \)

In practice, the fisher vector is defined only for the gradient of \( \mu \) and \( \sigma \).

Therefore: In practice, \( FV(X) \in R^{2D \cdot K} \)

\[
\frac{\partial L(\lambda|X)}{\partial \mu_{k,d}} = \sum_{i=1}^{n} T_{k,i} \cdot \frac{x_{i,d} - \mu_{k,d}}{\sigma_{k,d}^2}
\]

\[
\frac{\partial L(\lambda|X)}{\partial \sigma_{k,d}} = \sum_{i=1}^{n} T_{k,i} \cdot \left( \frac{(x_{i,d} - \mu_{k,d})^2}{\sigma_{k,d}^3} - \frac{1}{\sigma_{k,d}} \right)
\]

\[
p(z_i = k | x = x_i; \lambda) = T_{k,i} = \frac{\tau_{k} \cdot g(x; \mu_{k}, \sigma_{k})}{\sum_{r=1}^{K} \tau_{r} \cdot g(x; \mu_{r}, \sigma_{r})}
\]
Pooling methods – Bag of words – Vision

**Pros:** More information is preserved – “What kind of nose?”

With K clusters, BOW representation length is K and Fisher Vector is 2KD
Pooling methods – Bag of words – Vision

**Pros:** More information is preserved – “What kind of nose?”

Maybe they will have the same $T_{k,i}$ but their contribution to the fisher vector will be different:

\[
\frac{\partial L(\lambda|X)}{\partial \mu_{k,d}} = \sum_{i=1}^{n} T_{k,i} \cdot \frac{x_{i,d} - \mu_{k,d}}{\sigma_{k,d}^2}
\]

\[
\frac{\partial L(\lambda|X)}{\partial \sigma_{k,d}} = \sum_{i=1}^{n} T_{k,i} \cdot \left( \frac{(x_{i,d} - \mu_{k,d})^2}{\sigma_{k,d}^3} - \frac{1}{\sigma_{k,d}} \right)
\]
Hybrid Gaussian Laplacian Mixture Model

Why commit to a Gaussian distribution?

Univariate Gaussian: \( g(x, \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)

Univariate Laplacian: \( l(x, m, s) = \frac{1}{2s} e^{-\frac{|x-m|}{s}} \)

Univariate Hybrid Gaussian Laplacian: \( l(x, m, s)^b \cdot g(x, \mu, \sigma)^{1-b} \)
Hybrid Gaussian Laplacian Mixture Model

EM

Let $X_{trn} = \{x_1, x_2, ..., x_n\} \in R^D$ be the train set for the EM.

EM algorithm for HGLMM:
- Randomly initialize all the parameters $\lambda^{(0)} = \{\tau^0, \mu^{(0)}, \sigma^0, b^0, m^0, s^0\}$
- For $t=0...T$ (Or until convergence):
  - Estimation Step:
    \[
    p(z_i = k | x = x_i; \lambda^t) = T_{k,i}^t = \frac{\tau_k^t \cdot l(x, m_k^t, s_k^t)^{b_t} \cdot g(x, \mu_k^t, \sigma_k^t)^{1-b_t}}{\sum_{r=1}^{K} \tau_r^t \cdot l(x, m_r^t, s_r^t)^{b_t} \cdot g(x, \mu_r^t, \sigma_r^t)^{1-b_t}}
    \]
Hybrid Gaussian Laplacian Mixture Model

EM

Maximization Step:

\[
T_{k}^{(t+1)} = \frac{\sum_{i=1}^{N} T_{k,i}^{(t)}}{\sum_{r=1}^{K} \sum_{i=1}^{N} T_{r,i}^{(t)}}
\]

\[
\mu_{k,d}^{(t+1)} = \frac{\sum_{i=1}^{N} T_{k,i}^{(t)} \cdot x_{i,d}}{\sum_{i=1}^{N} T_{k,i}^{(t)}}
\]

\[
(\sigma_{k,d}^{(t+1)})^2 = \frac{\sum_{i=1}^{N} T_{k,i}^{(t)} (x_{i,d} - \mu_{k,d}^{(t+1)})^2}{\sum_{i=1}^{N} T_{k,i}^{(t)}}
\]

\[
\sum_{m_{k,d}^{(t+1)} \leq x_{i,d}} T_{k,i}^{(t)} = \sum_{m_{k,d}^{(t+1)} > x_{i,d}} T_{k,i}^{(t)}
\]

\[
\delta_{k,d}^{(t+1)} = \frac{\sum_{i=1}^{N} T_{k,i}^{(t)} |x_{i,d} - m_{k,d}^{(t+1)}|}{\sum_{i=1}^{N} T_{k,i}^{(t)}}
\]
Hybrid Gaussian Laplacian Mixture Model

EM

\[ L_{b_k,d} = \sum_{i=1}^{N} T_{k,i}^{(t)} \left( -\log (2s_{k,d}) - \frac{|x_{i,d} - m_{k,d}|}{s_{k,d}} \right) \]

\[ G_{b_k,d} = \sum_{i=1}^{N} T_{k,i}^{(t)} \left( -\log \left( \sqrt{2\pi\sigma_{k,d}} \right) - \frac{(x_{i,d} - \mu_{k,d})^2}{2\sigma_{k,d}^2} \right) \]

Then the contribution of \( b_{k,d} \) to the log-likelihood is:

\[ b_{k,d} \cdot L_{b_k,d} + (1 - b_{k,d}) \cdot G_{b_k,d} \]
Therefore, under the constraint that $0 \leq b_{k,d} \leq 1$, the value of $b_{k,d}^{(t+1)}$ that maximizes the log-likelihood is:

$$b_{k,d}^{(t+1)} = \begin{cases} 1 & \text{if } L_{b_{k,d}}^{(t+1)} > G_{b_{k,d}}^{(t+1)} \\ 0 & \text{otherwise} \end{cases} \quad (3.9)$$
Hybrid Gaussian Laplacian Mixture Model Fisher Vector

Hence, there is hard selection for each coordinate and each component

For $b_{k,d} = 0$:

$$\frac{\partial \mathcal{L}(X|\lambda)}{\partial \mu_{k,d}} = \sum_{i=1}^{N} T_{k,i} \cdot \frac{x_{i,d} - \mu_{k,d}}{\sigma_{k,d}^2}$$

$$\frac{\partial \mathcal{L}(X|\lambda)}{\partial \sigma_{k,d}} = \sum_{i=1}^{N} T_{k,i} \left( \frac{(x_{i,d} - \mu_{k,d})^2}{\sigma_{k,d}^3} - \frac{1}{\sigma_{j,d}} \right)$$

For $b_{k,d} = 1$:

$$\frac{\partial \mathcal{L}(X|\lambda)}{\partial m_{k,d}} = \sum_{i=1}^{N} \frac{T_{k,i}}{s_{k,d}} \cdot \left\{ \begin{array}{ll} 1 & \text{if } x_{i,d} > m_{k,d} \\ -1 & \text{otherwise} \end{array} \right.$$  

$$\frac{\partial \mathcal{L}(X|\lambda)}{\partial s_{k,d}} = \sum_{i=1}^{N} T_{k,i} \left( \frac{|x_{i,d} - m_{k,d}|}{s_{k,d}^2} - \frac{1}{s_{k,d}} \right)$$
Fisher Information Matrix (FIM)

More algebra (14 page supplementary)
**Sentence representation**

A man is playing guitar

Local Feature Extraction

\[
\text{word2vec(“A”) = (131, 128, 111, 10, \ldots, 14, 11) } \in \mathbb{R}^D
\]

\[
\text{word2vec(“playing”) = (11, 61, 2, 13, \ldots, 11, 10) } \in \mathbb{R}^D
\]

\[
\text{word2vec(“guitar”) = (21, 122, 14, 1, \ldots, 110, 1) } \in \mathbb{R}^D
\]
**Sentence representation**

\[
\text{word2vec}("A") = (131, 128, 111, 10, ..., 14, 11) \in \mathbb{R}^D \\
\text{word2vec}("playing") = (11, 61, 2, 13, ..., 11, 10) \in \mathbb{R}^D \\
\text{word2vec}("guitar") = (21, 122, 14, 1, ..., 110, 1) \in \mathbb{R}^D
\]

\[\text{Pooling} \quad HGLMM\ FV \in \mathbb{R}^m\]

**Generic representation:** dataset independent
Bringing Images and Sentences to the same domain

Given a training set of images and their matching sentences: \{\langle \text{Image}, \text{Sentence} \rangle \}

We present each image by the VGG(Image) representation. \( VGG(\text{Image}) \in R^{D_{\text{Image}}} \)

We present each sentence by the HGLMM Fisher Vector representation on top of the word2vecs of the words in the sentence. \( FV(\text{Sentence}) \in R^{D_{\text{Sentence}}} \)

We apply the **Canonical Correlation Analysis** on the training set, which as a result returns two projection matrices, \( W_{\text{Image}} \) and \( W_{\text{Sentence}} \), such that:

\[
W_{\text{Image}} \cdot VGG(\text{Image}) \in R^{\text{common}}
\]

\[
W_{\text{Sentence}} \cdot FV(\text{Sentence}) \in R^{\text{common}}
\]
Image Annotation

A man is playing guitar
Two girls are playing soccer
A man is walking down the street
A man is climbing a mountain
A boy is riding a bicycle
Image Annotation
Image Annotation
Sentence Synthesis

$W_0 = \text{"two"}$

LSTM

CCA(I)

CNN(I)
Sentence Synthesis

\[ W_0 = \text{"two"} \quad W_1 = \text{"girls"} \]

\[
\begin{align*}
\text{LSTM} & \rightarrow \text{CCA}(W_0) \\
\text{CCAI} & \rightarrow \text{CNN}(I) \\
\text{HGLMM}(W_0) & \rightarrow \text{word2vec}(W_0)
\end{align*}
\]
Sentence Synthesis

$W_0 =$ "two"

$W_1 =$ "girls"

$W_2 =$ "playing"

LSTM

CCA(I)

CNN(I)

LSTM

CCA(W_0)

HGLMM(W_0)

word2vec(W_0)

LSTM

CCA(W_1)

HGLMM(W_1)

word2vec(W_1)
Sentence Synthesis

\[
\begin{align*}
W_0 = "two" & \quad W_1 = "girls" & \quad W_2 = "playing" & \quad W_3 = "soccer" & \quad \text{ENDSENT} \\
\text{LSTM} & \quad \text{LSTM} & \quad \text{LSTM} & \quad \text{LSTM} & \quad \text{LSTM} \\
\text{CCA(I)} & \quad \text{CCA}(W_0) & \quad \text{CCA}(W_1) & \quad \text{CCA}(W_2) & \quad \text{CCA}(W_3) \\
\text{CNN(I)} & \quad \text{HGLMM}(W_0) & \quad \text{HGLMM}(W_1) & \quad \text{HGLMM}(W_2) & \quad \text{HGLMM}(W_3) \\
\text{word2vec}(W_0) & \quad \text{word2vec}(W_1) & \quad \text{word2vec}(W_2) & \quad \text{word2vec}(W_3) \\
\end{align*}
\]
Results

Synthesis:
94% of the generated sentences are new compared to 20% in Google’s system

imageannotator.cs.tau.ac.il
a dog with ball in its mouth
a basketball player in the uniform is running in the air
a man in black helmet is riding bike on the road
a boy is jumping into pool
The model was trained on only 8,000 images. Two dogs are playing in the water.
a man in red shirt is riding his bike

a dog is jumping up at the ground

a man is sitting on the ground

a man is holding up his hand on the ground
<table>
<thead>
<tr>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /> <img src="image2.png" alt="Image 2" /> <img src="image3.png" alt="Image 3" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correct sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>A brunette woman taking a big bite of some food.</td>
</tr>
<tr>
<td>A young woman in black enjoys a bite to eat.</td>
</tr>
<tr>
<td>A woman in a black top mugs for the camera.</td>
</tr>
<tr>
<td>A woman wearing a black shirt is eating.</td>
</tr>
<tr>
<td>A woman about to eat in a white room.</td>
</tr>
<tr>
<td>A black dog walking on the beach after swimming in the ocean.</td>
</tr>
<tr>
<td>A black dog walks on the beach near the rocks.</td>
</tr>
<tr>
<td>A black dog walks along an ocean front.</td>
</tr>
<tr>
<td>A black dog walks on the sand.</td>
</tr>
<tr>
<td>A black dog on a rocky beach.</td>
</tr>
<tr>
<td>Two women dressed for cold weather in jackets and gloves look at something on a</td>
</tr>
<tr>
<td>cellphone screen.</td>
</tr>
<tr>
<td>Two teen girls are looking at a small electronic device while wearing winter</td>
</tr>
<tr>
<td>coats.</td>
</tr>
<tr>
<td>Two girls holding drinks and looking at something on a cellphone.</td>
</tr>
<tr>
<td>Two girls bundled up in winter coats pose for a picture.</td>
</tr>
<tr>
<td>Two young girls huddle to look at a cellphone.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Our five nearest neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>A woman wearing a black shirt is eating.</td>
</tr>
<tr>
<td>A brunette woman taking a big bite of some food.</td>
</tr>
<tr>
<td>Two girls holding drinks and looking at something on a cellphone.</td>
</tr>
<tr>
<td>Girls are preparing food to eat.</td>
</tr>
<tr>
<td>A young woman in black enjoys a bite to eat.</td>
</tr>
<tr>
<td>Asian girls serving food on trays.</td>
</tr>
<tr>
<td>A black dog on a rocky beach.</td>
</tr>
<tr>
<td>A black dog walks on the beach near the rocks.</td>
</tr>
<tr>
<td>A black dog walks on the sand.</td>
</tr>
<tr>
<td>A black dog walks along an ocean front.</td>
</tr>
<tr>
<td>Three dogs run on beach, two playing with unknown object.</td>
</tr>
<tr>
<td>Two girls holding drinks and looking at something on a cellphone.</td>
</tr>
<tr>
<td>Two women sharing a drink taking a blurred photo of their faces</td>
</tr>
<tr>
<td>A man has is arm around the woman who is holding a metallic object up to her face.</td>
</tr>
<tr>
<td>A woman wears a scarf and uses her smartphone while seated</td>
</tr>
<tr>
<td>A woman in a red coat drinking coffee.</td>
</tr>
</tbody>
</table>
Results

**Correct sentences**

- A brunette woman taking a big bite of some food.
- A young woman in black enjoys a bite to eat.
- A woman in a black top mug for the camera.
- A woman wearing a black shirt is eating.
- A woman wearing a black shirt is eating.
- A black dog walking on the beach after swimming in the ocean.
- A black dog walks along an ocean front.
- A black dog walks on the sand.
- A black dog on a rocky beach.
- Two women dressed for cold weather in jackets and gloves look at something on a cellphone screen.
- Two teen girls are looking at a small electronic device while wearing winter coats.
- Two girls holding drinks and looking at something on a cellphone.
- Two girls huddled up in winter coats pose for a picture.
- Two young girls huddle to look at a cellphone.
- A red liquid travels down a grate as it leaks across a street.
- A person crosses the street.
- A closeup of a person walking across the street.
- An oil stained sidewalk with a walker next to it.
- A person is crossing the street.
- Two men sitting in front of a sign that says game preserve while surrounded by dead pheasants.
- Two men sitting in front of an "Etohah Valley Game Preserve" sign.
- Two men in wife beaters sit in front of a bunch of dead birds.
- Two young men are sitting in front of a dozen bowls.
- A man with yellow shoes and a gray hat sits on an orange and blue teeter-totter with a woman.
- A man in a sweater is on the left side of a seesaw with a woman in blue on the right side.
- An adult male and female balance on a seesaw.
- Two people are playing on a teeter totter.
- A man and a woman are on a seesaw.

**Our five nearest neighbors**

- A woman wearing a black shirt is eating.
- A brunette woman taking a big bite of some food.
- Two girls holding drinks and looking at something on a cellphone.
- Girls are preparing food to eat.
- A young woman in black enjoys a bite to eat.
- Asian girls serving food on trays.
- Two women standing near a sign.
- A woman is holding a picket sign on a sidewalk.
- Men and women in a crowd are holding signs.
- Large group of people with 2 women in the front, one in a yellow shirt and a skirt and the other in a red shirt and white shorts.
- Many people walking, some are holding signs.
- A young boy poses for a picture in front of a playground.
Image annotation

Flickr 8K
Flickr 30K
Coco
Pascal 1K

Bengio, Toronto+Google
Zemel, Toronto
Yuille, Baidu+UCLA
Socher, Stanford+Google
Fei-Fei, Stanford
Our, Blavatnik School of CS
Image search

- Flickr 8K
- Flickr 30K
- Coco
- Pascal 1K

- Bengio, Toronto+Google
- Zemel, Toronto
- Yuille, Baidu+UCLA
- Socher, Stanford+Google
- Fei-Fei, Stanford
- Our, Blavatnik School of CS
Future work: image Q&A

Input: image + a question

What are the girls doing?

Output: the answer

They are playing soccer.
Summary

Task I: Image Annotation
- Two girls are playing soccer
- A man is walking down the street
- A man is climbing a mountain
- A boy is riding a bicycle

Task II: Image Search
- A man is playing guitar

Task III: Description Synthesis
- Input: new unseen image
- Output: new description in English

Two girls playing soccer.

Image representation | Text representation

Hybrid Gaussian Laplacian Mixture Model Fisher Vector

Sentence Synthesis

We prove that there is hard selection for each coordination and each component.

For $\beta_{kd} = 0$:

$$W_{kd}(x) = \sum_{s=1}^{S} \frac{1}{\sigma_s^2} \left(1 - \frac{s}{\sigma_s^2} \right) r_{s,k}^d$$

For $\beta_{kd} = 1$:

$$W_{kd}(x) = \sum_{s=1}^{S} \frac{1}{\sigma_s^2} \left(1 - \frac{s}{\sigma_s^2} \right) r_{s,k}^d$$